applied sciences

MDPI

Article

OISE: Optimized Input Sampling Explanation with a Saliency Map Based on the Black-Box Model

Zhan Wang and Inwhee Joe *

Computer Science, Hanyang University, Seoul 04763, Republic of Korea; zhanbaobao6@gmail.com
* Correspondence: iwjoe@hanyang.ac.kr

Abstract: With the development of artificial intelligence technology, machine learning models are becoming more complex and accurate. However, the explainability of the models is decreasing, and much of the decision process is still unclear and difficult to explain to users. Therefore, we now often use Explainable Artificial Intelligence (XAI) techniques to make models transparent and explainable. For an image, the ability to recognize its content is one of the major contributions of XAI techniques to image recognition. Visual methods for describing classification decisions within an image are usually expressed in terms of salience to indicate the importance of each pixel. In some approaches, explainability is achieved by deforming and integrating white-box models, which limits the use of specific network architectures. Therefore, in contrast to white-box model-based approaches that use weights or other internal network states to estimate pixel saliency, we propose the Optimized Input Sampling Explanation (OISE) technique based on black-box models. OISE uses masks to generate saliency maps that reflect the importance of each pixel to the model predictions, and employs black-box models to empirically infer the importance of each pixel. We evaluate our method using deleted/inserted pixels, and extensive experiments on several basic datasets show that OISE achieves better visual performance and fairness in explaining the decision process compared to the performance of other methods. This approach makes the decision process clearly visible, makes the model transparent and explainable, and serves to explain it to users.

Keywords: XAI; black-box model; mask; saliency map; importance; explanation

check for updates

Citation: Wang, Z.; Joe, I. OISE: Optimized Input Sampling Explanation with a Saliency Map Based on the Black-Box Model. *Appl. Sci.* **2023**, *13*, 5886. https://doi.org/10.3390/app13105886

Academic Editor: João M. F. Rodrigues

Received: 29 March 2023
Revised: 28 April 2023
Accepted: 5 May 2023
Published: 10 May 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, artificial intelligence has been applied in many fields and is widely recognized for its excellent results in many traditionally complex problems. Currently, artificial intelligence (AI) techniques have demonstrated near-human-level performance in a variety of computer vision operations, such as image classification and object perception [1], and have been successfully applied. The currently used models have excellent performance in terms of accuracy, but the models do not provide proper descriptions and act as a black box. With the development of AI technology, machine learning models are becoming increasingly complex, and the accuracy of the models is improving, but the transparency of the models is decreasing, and a significant part of the decision process of both is still unclear and difficult to explain to users, making it impossible to fully understand the function or logic behind it. This feature is considered to be one of the biggest problems in the application of AI technology. As black-box machine learning models are increasingly used to make important predictions in critical contexts, the demand for transparency from various stakeholders in AI is growing [2]. The opacity of machine decision making reduces human trust in artificial intelligence. For example, in the medical field, a deep learning system determines whether a patient has cancer based on medical images [3], while a human medical expert has the opposite opinion. Since the system cannot provide an explanation, the expert may not accept the system's opinion, and if the judgment is wrong, it may lead to medical errors. No matter what kind of application, we can see that if we

do not solve the problem of explainability of deep learning, the future development of its application will be limited.

With the development of intelligent systems in application areas, such as autonomous robots and vehicles [4–6], health care [7–9], such as soft tissue sacromas segmentation [10], skin lesion segmentation [11], and coronavirus (COVID-19) classification [12], classification and detection in image processing [13–15], etc. Automated systems must provide users, developers, and regulators with explanations based on practical factors and social and legal reasons when making decisions or recommendations. Therefore, making the black-box model transparent and explainable is also an important research in the field of AI. The technique for making models transparent and explainable is called XAI, and XAI techniques improve the reliability of models by giving users confidence that the models make good decisions [16]. Many existing methods compute the importance of a given base model and output class [17–19], but they require the use of intermediate feature mappings, network weights, and other internal factors of the underlying model. The explainability of the AI model is shown in Figure 1.

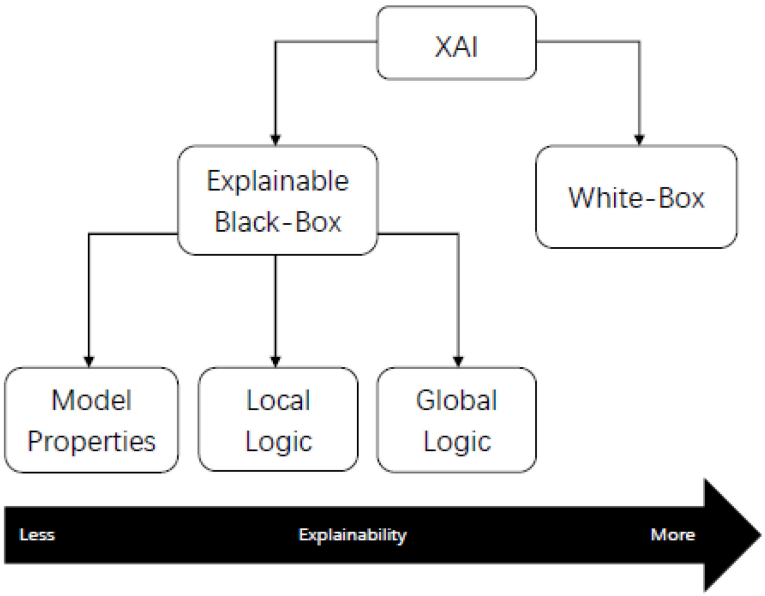


Figure 1. Explainability of the AI model. Some methods require explainability through deformation and integration of white-box models, which limits the use of specific network architectures. Therefore, instead of white-box approaches that use gradients or other internal network states to estimate pixel importance, techniques that use black-box models are proposed.

A black-box model can be interpreted as follows:

(1) Model properties: Presentation of specific properties of the model or its predictions, such as sensitivity to changes in the properties or identification of the model components responsible for a given decision.
(2) Local logic: Presentation of the internal logic behind an individual decision or prediction.
(3) Global logic: Representation of all internal logic.

The black-box model problem specifically involves the following three points [20]:

(1) Inability to dig causality problem: The internal structure of the black-box model is complicated, and when making predictions, we will evaluate the goodness of the model based on some model evaluation indicators (such as AUC), but even if the AUC is high, it is still unclear whether the black-box model is based on the correct judgment. If the model cannot provide a reasonable causal relationship, the results of the model will be difficult to be convincing.
(2) Insecurity problem of black-box model: For modelers, the internal structure of black-box models is complicated, and it is usually difficult to detect these attacks when the

models are attacked from the outside; for users of the models, they do not understand the operating mechanism of the models and only use the results of the models to make decisions. It is difficult for users to detect anomalies from the results of the black-box model, which may cause the problem of insecurity in the use of the model results.

(3) Possible bias problem in the black-box model: When making predictions, it reinforces the problem of data imbalance that may exist in the data collection process, which leads the model to end up with biased results.

This is also a problem that needs to be overcome in the future.

Here, we propose a new black-box approach for estimating pixel saliency. By inserting and removing pixels to estimate the weights corresponding to different pixels, and visualizing the saliency range, the saliency of different pixel points is presented in the form of heat maps to highlight the key pixels for the purpose of explanatory illustrations to humans. With OISE, we can clearly see which region of the image the network is focused on, and improve its inability to classify multiple targets in the same image compared to traditional deep learning explainable methods. Unlike traditional CAM series methods that require changing the network structure, OISE does not require internal access to arbitrary networks, does not require reimplementation of each network architecture, applies to existing image networks, and is considered a complete black box with no assumptions about parameters, features, and classification.

Our contributions are as follows:

(1) We have improved the Randomized Input Sampling for Explanation (RISE) method by using an optimized way to generate the mask, which reduces the computational effort and makes the generated range of significance regions more accurate.

(2) We introduce a new black-box resolution method that compensates for the shortcomings of perturbation-based, intuitive, and understandable representation of the weighted value of activity.

(3) We evaluate the generation. The saliency map shows that the fairness of the work can be identified, and points out that this method can find better evidence for the target category.

In Section 2, we introduce the related works and summarize the shortcomings of the existing method and improve it accordingly; in Section 3, we give a detailed description of the implementation process of the OISE method; in Section 4, the experimental process and results are described, and the experimental results are evaluated using the pixel deletion and insertion methods to confirm the practicality and accuracy of the OISE method. Finally, Section 5 summarizes the main ideas of OISE and discusses the advantages and disadvantages of the method, application areas, and future research directions.

2. Related Works

Researchers have explored many directions in the field of explainable artificial intelligence, and the importance of interpretation has been widely studied in various fields of machine learning and deep learning.

The Randomized Input Sampling for Explanation (RISE) method [21] introduced by Petsiuk et al. perturbs an input image by multiplying it with randomized masks. RISE uses the black-box model, which differs from the white-box method, which uses other internal network states to infer pixel importance. The black-box model uses mask-based visualization by estimating the importance of the input image region for the model prediction. The method generates the mask by first sampling the smaller binary mask and then upsampling it using bilinear interpolation to improve resolution. After interpolation, the mask Mi is no longer binary, but has the value [0, 1]. To make the mask more flexible, all masks are shifted by a random number of pixels in both spatial directions. The saliency of the pixels is then estimated by randomly combining dimmed pixels to reduce their intensity to zero, and this model is built by multiplying the image with a mask of [0, 1] values. Saliency maps are generated by empirically estimating the sum using Monte Carlo

sampling. This black-box based interpretation method generates multiple masks by random or Monte Carlo sampling to compute the saliency of each mask field, which usually requires a lot of masks and computations. It is very complex and wastes time and resources.

To localize visual evidence in images, Class Activation Mapping (CAM) [22] emerged in 2016. In CAM, the authors argue that the global average pooling layer has local localization capability, replace the original pooling and fully connected layers after the convolutional network with global average pooling and fully connected layers, retrain the training model to obtain the weights, and obtain the deep feature maps' weighted sum to build the saliency map. The class activation map is simply a weighted linear sum of the presence of these visual patterns at different spatial locations. By simply upsampling the class activation map to the size of the input image, the image regions most relevant to a particular class can be identified, providing a new idea for the explainability of convolutional neural networks. CAM can also be used in many other ways. However, this method can only be applied to a specific CNN architecture, and the importance of each feature map is represented by retraining the model to obtain the corresponding weights on the fully connected layer. This technique is very useful, but has some drawbacks: first, it requires changing the network structure, for example, by changing the fully connected layer to a globally averaged pooling layer, which does not facilitate training; second, it is a visualization technique based on a classification problem, which is not as effective for regression problems.

To address the shortcomings of CAM, an improved technique, Gradient-weighted Class Activation Mapping (Grad-CAM) [23] , emerged in 2017, which allows visualization without changing the network structure. Grad-CAM extends CAM by weighting the feature activation value for each position and the class average weight for each feature mapping channel. First, given an image and a target class as input, the image is propagated through the CNN part of the model, and then the raw score for that class is obtained by task-specific computation. For all classes, the gradient is set to zero, except for the gradient of the target class, which is set to one. This signal is then backpropagated to rectified convolutional feature maps, which are combined to compute the rough Grad-CAM localization (blue heat map), which indicates where the model needs to make precise decisions. Finally, the heat map is multiplied point-by-point with Guided Backprop to obtain a high-resolution and semantically specific visualization of Guided Grad-CAM. The difference between this and CAM is that Grad-CAM adds a ReLU to the final weighted sum, the reason being that we only care about pixel points that have a positive impact on the target class, and without the ReLU layer, we may end up bringing in some pixels that belong to other classes, thus affecting the interpretation.

Haofan Wang, Zifan Wang et al. proposed Score-CAM [24], which follows the main idea of CAM (linear weighting of the feature map). Compared with the previous series of CAM methods, the main difference is the way to obtain the linear weights. The first generation of CAM used the model weights on the full connection layer after training. Grad CAM and Grad-CAM++ [25] both used the local gradients on the corresponding feature map (the difference is in the method of processing the gradients). Unlike previous methods based on class activation maps, Score-CAM obtains the weights of each activation map by forward-passing the scores of each activation map on the target class, thus eliminating the dependence on gradients, and the final result is obtained by a linear combination of weights and activation maps. The results of the study show that Score-CAM has better visualization and fairness in explaining the decision process. Score-CAM not only locates a single object accurately, but also shows better performance than previous work in locating multiple objects of the same type. Grad-CAM tends to capture only one target in the image, and both Grad-CAM++ and Score-CAM show the ability to locate multiple targets; however, Score-CAM's remarkable map is more focused than Grad-CAM++.

Marco Tulio Ribeiro et al. proposed the locally interpretable model diagnostic interpretation (LIME) in 2016 [26]. LIME is an algorithm that interprets the predictions of a classifier or regressor by performing a local approximation with an interpretable model. It

modifies a single data sample by adjusting the feature values and observing the effect on the output. The output of LIME is a set of interpretations representing the contribution of each feature to the prediction of a single sample, which is a local interpretability. However, the LIME algorithm is slow, and the results must be predicted once by the original model for each sampled image after sampling is complete.

In summary, the RISE method generates multiple masks by random or Monte Carlo sampling to calculate the importance of each mask field, which usually requires a lot of masks and calculations. It is very complex and wastes time and resources. The existing CAM series methods have been able to interpret image targets more accurately in terms of localization, but the evaluation metrics need to be artificially estimated and the model architecture needs to be changed, which cannot locate multiple objects or locate them inaccurately. The LIME algorithm has strong generality and does not need to change the model internals, but the result is a local approximation of the black box model instead of a global approximation, and when the input is perturbed, the samples obey the Gaussian distribution, ignore the correlation between features, and is not stable enough to obtain completely different results for repeated interpretations using the same parameters and the same method.

3. Proposed Method

3.1. Framework of OISE

OISE is a new black-box interpretation method for better visual interpretation. By optimizing the loss function and continuously updating the mask, we can reduce the computation of the random generation mask, minimize the area of the generation mask and influence the decision score, and make the generated saliency map area more accurate. The model framework is shown in Figure 2.

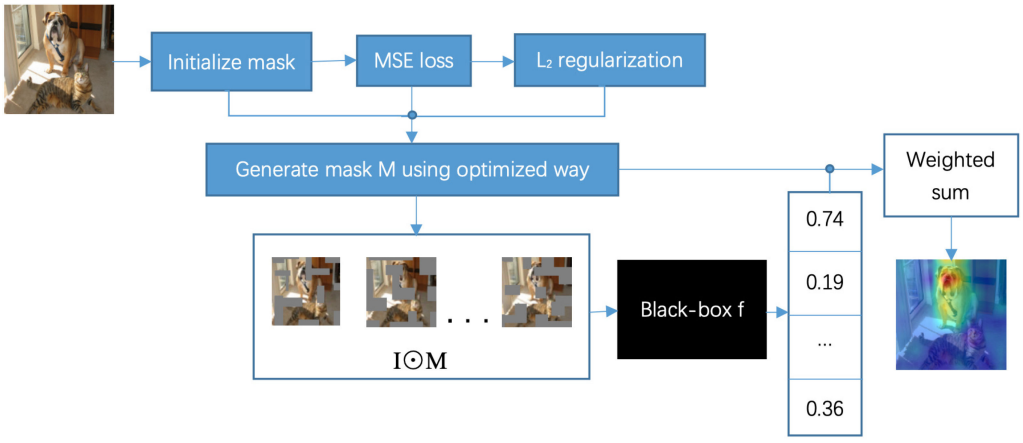


Figure 2. Flowchart of the OISE algorithm. This is the flowchart of the OISE technology. The masks are generated in an optimal way and the masks are continuously updated by optimizing the loss function to minimize the generated mask area and influence the decision score.

3.2. Definition of the OISE-Based Black-Box Method

Using random or Monte Carlo sampling to generate multiple masks and compute the significance of each mask typically requires a large number of masks and computations. It is complex, wastes time and resources, and the range of significant features in the generated saliency map is not particularly accurate.

The OISE method can generate a saliency map without accessing a network, and does not need to rebuild the network architecture. It is applicable to all image meshes.

The mask is generated in an optimized way, the loss function is optimized, and the mask is continuously updated to minimize the generated mask area, which affects the decision score. When computing the mask, we should reduce the amount of computation, reduce the complexity, and reduce the key display area of the saliency map in the final

output. The main content is to subsample the input image with the mask, record the response to each masked image, and then detect the basic model. The weights are derived from the output probabilities predicted by the masked images in the base model. The linear combination of multiplying the weights and masks and then adding them together is the final saliency map.

$$S_{I,f}(\lambda) \approx \frac{1}{E[M] \cdot N} \sum_{i=1}^{N} f(I \odot M_i) \cdot M_i(\lambda) \tag{1}$$

where f is a black-box model that produces scalar confidence scores for a given input of I, I is the input image, m is the binary mask, and $I \odot M$ denotes element multiplication. The importance of a pixel λ is its expected score on the mask M. The more important the pixel, the higher $I \odot M$ is.

First, input the image I to generate a mask pointing to the input image.

Mask generation process:

(1) Initialize a random mask.

(2) Use gradient descent technology to optimize the MSE (Mean Square Error) loss function. X, Y are the horizontal and vertical coordinates of the input image. To learn the desired model, we find the optimal parameters by minimizing the cost function θ.

$$MSE = \frac{1}{N}(Y - X\theta)^T(Y - X\theta) \tag{2}$$

Choose the initialized parameter value. For example, $\theta = (0, 0, 0, \ldots, 0)$. Then, select the step $\alpha = 0.1$. Calculate the partial differential according to the loss function:

$$\nabla MSE(\theta) = \begin{bmatrix} \frac{dMSE(\theta)}{d\theta_0} \\ \frac{dMSE(\theta)}{d\theta_1} \\ \vdots \\ \frac{dMSE(\theta)}{d\theta_p} \end{bmatrix} = \frac{2}{N} X^T(X\theta - Y) \tag{3}$$

Update θ until convergence.

$$\theta = \theta - \alpha \nabla MSE(\theta) = \theta - \alpha X^T(X\theta - Y) \tag{4}$$

(3) Add L_2 regulation to make the parameters close to 0, but not equal to 0. Reducing the parameter size, complexity, and mask area will affect the decision score:

$$loss_{L_2}(w) = \sum_i w_i^2 \tag{5}$$

(4) Update mask M according to the optimized loss function to mask the input image I.

Then, the input image I is multiplied by the mask M_i to obtain the masked image $I \odot M_i$ with i = 1, …, N.

Input the cover image in the basic model f and output the weight value.

The weights are the probability scores generated by the masks and are adjusted according to the distribution of the masks. The final saliency map is generated by a linear combination of the weights and the masked images, multiplied, and then summed.

Finally, the weighted value of the mask is taken to obtain the saliency map. The complete process is shown in Algorithm 1.

Algorithm 1 OISE

Input: Image I, mask M, model f
Output: I's saliency map (linear combination of masks)

for i <= N **do**
    $Y \Leftarrow 0.1X + 0.2$
    MSE loss
    L_2 regularization
    $\alpha \Leftarrow 0.1$
    $M \Leftarrow$ optimized mask
    θ close to 0
end for
masked image $\leftarrow I \odot M$
$w \leftarrow$ output probability of the masked image prediction in the f
saliency map

Time complexity: O(N)

4. Experimental Results and Discussion

4.1. Experimental Results

In this section, we conduct experiments to evaluate the effectiveness of the proposed interpretation methods. First, we qualitatively evaluated our method using ImageNet visualization to demonstrate the effectiveness of class conditional location of objects in a given image. In our experiment, we used the publicly available object classification dataset, i.e., ILSVRC2012 and PASCAL VOC 2007. We used H = W = 224 throughout.

The ILSVRC2012_img_val dataset from ImageNet contains 50,000 images, 50 of each type. These categories correspond to the set of 1000 synonyms in WordNet. If an image contains x, it belongs to category x, where x is a synonym. The PASCAL VOC 2007 standard dataset is a benchmark for measuring the ability to classify images. The dataset contains 5011 images in the training set and 4952 images in the test set, for a total of 9963 images with 20 categories.

The evaluation was performed for the top-1 and top-5 predicted categories, and 5000 images were selected from the dataset for evaluation. Given an image, we first obtained category predictions from our network and then generated OISE saliency maps for each predicted category. We used the pre-trained VGG-16 [27], GoogleNet [28], and Resnet50 [29] to evaluate OISE. After evaluating ILSVRC2012 and PASCAL VOC 2007, we report the values set in the localization error table for the top-1 and top-5 rankings in Table 1. In all three classical neural networks, OISE has a lower localization error than CAM and Grad-CAM. CAM and Grad-CAM require changes in the model structure and must be retrained, resulting in a worse classification error, while OISE improves in classification performance.

The accuracy is shown in Table 2 for the ILSVRC2012 and PASCAL VOC07 datasets. OISE performs with consistently high accuracy.

As shown in Table 3, OISE has an average decrease rate of 47.4% and an average increase rate of 19.7%, which is better than other CAM-based methods. The original input is masked by point-wise multiplication with the saliency maps to observe the score change on the target class. We follow the metrics used in [25] to measure the quality, the average drop is expressed as $\sum_{i=1}^{N} \frac{max(0, Y_i^c - O_i^c)}{Y_i^c} \times 100$ and the average increase is expressed as $\sum_{i=1}^{N} \frac{Sign(Y_i^c < O_i^c)}{N}$, where Y_i^c is the predicted score for class c on image im and O_i^c is the predicted score for class c with the explanation map region as input. The experiment is performed on the ILSVRC2012 validation set; 2000 images are randomly selected.

OISE performs well in the recognition task and can successfully detect distinguishable regions of the target objects. The results of the recognition task show that OISE better reflects the decision process of the original CNN model than previous methods. The statistical graph of the comparison results is shown in Figure 3.

Table 1. ILSVRC2012 Val classification and positioning error (%) of VGG-16, GoogleNet, and Resnet50 (the lower the better). OISE achieves excellent positioning error without compromising classification performance. Except for OISE, all of them are white-box models.

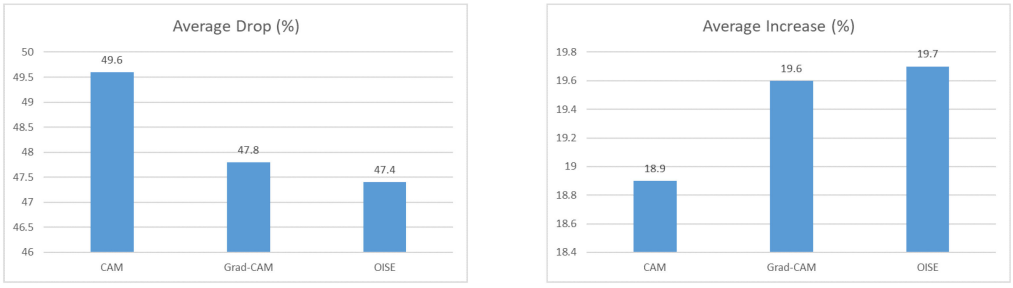
Model	Method	Classification		Localization	
		Top-1	Top-5	Top-1	Top-5
VGG-16	CAM	33.40	12.20	57.20	45.14
	Grad-CAM	30.38	10.89	56.51	46.41
	OISE	30.31	10.23	56.82	46.38
GoogleNet	CAM	31.9	11.3	60.09	49.34
	Grad-CAM	31.9	11.3	60.09	49.34
	OISE	30.8	10.6	60.89	49.12
Resnet50	CAM	32.4	11.9	58.06	48.62
	Grad-CAM	31.7	10.6	57.62	47.28
	OISE	30.6	10.1	56.74	46.32

Table 2. Mean accuracy (%) of ILSVRC2012 and PASCAL VOC 2007 on different models. Except for OISE, all of them require white-box models.

Dataset	Modle	Accuracy
ILSVRC2012	VGG-16	75.26
	Resnet50	84.43
PASCAL VOC 2007	VGG-16	78.64
	Resnet50	86.54

Table 3. Evaluation results on recognition (lower is better in Average Decrease, higher is better in Average Increase).

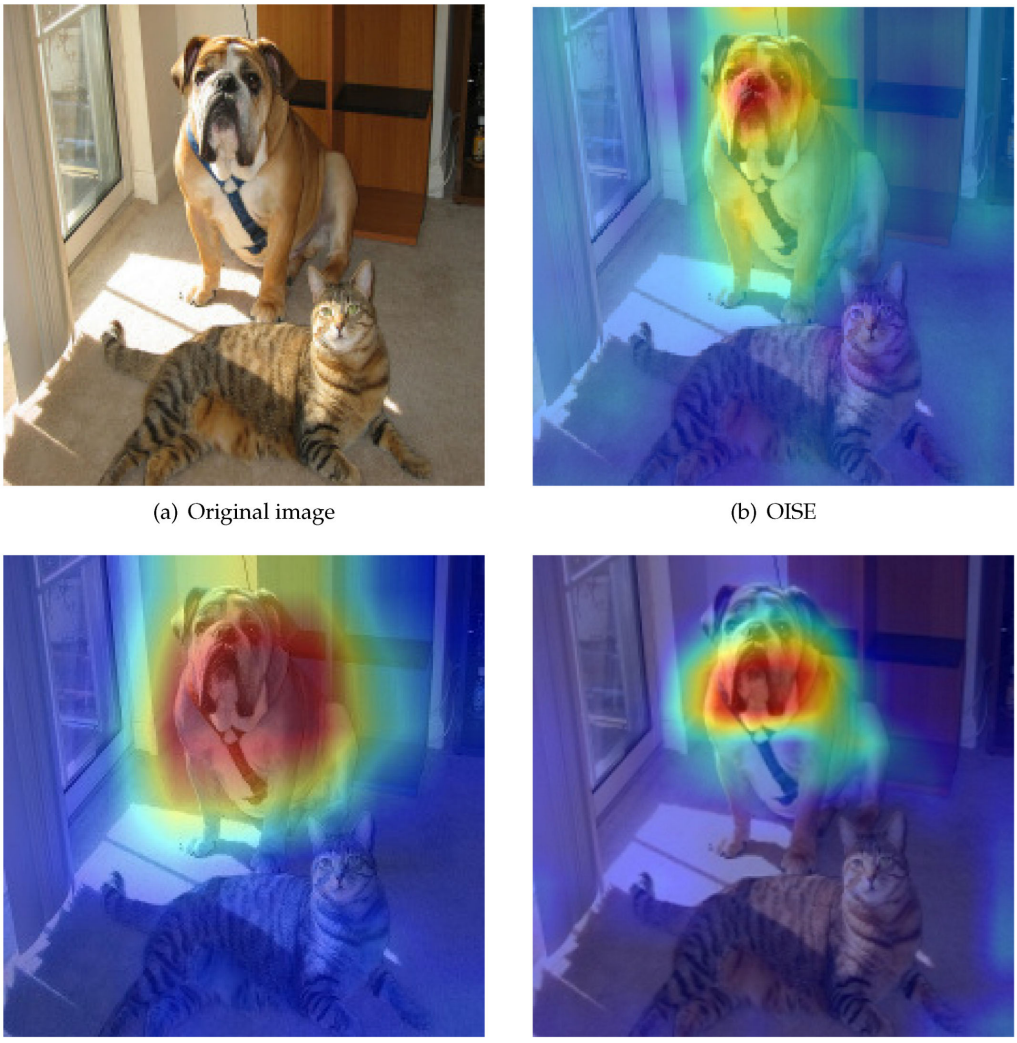
Method	CAM	Grad-CAM	OISE
Average Decrease (%)	49.6	47.8	47.4
Average Increase (%)	18.9	19.6	19.7



(a) Average Decrease (%) (b) Average Increase (%)

Figure 3. Average decrease and increase.

We qualitatively compared the saliency maps generated by the three methods (CAM, Grad-CAM, and OISE). Our method produces a more intuitively explainable saliency map with less random noise. The results are shown in Figure 4.



(a) Original image (b) OISE

(c) CAM (d) Grad-CAM

Figure 4. Identification results. (**a**) The original image. (**b**) The result of using the OISE algorithm. (**c**) The result of using CAM. (**d**) The result of using Grad-CAM. The visual description of using OISE gives a good description in terms of importance. The reason why the image was identified as a bulldog was pointed out.

The method can explain the reason to people in a clearer way, and is more convenient compared to the white-box method, for arbitrary networks, without internal access, without reimplementing each network architecture, and for existing image networks, and is considered to be completely black-box. This method provides a more precise positioning capability than traditional CAM series methods. Grad-CAM is unable to perform multi-target detection, and Score-CAM improves on Grad-CAM's shortcomings with more accurate localization and a poor classification task. OISE has excellent performance in both localization and classification tasks. However, the method also has some limitations, such as that it requires a lot of computation when generating and updating the mask, and the optimal parameters for optimizing the mask may be different for each update due to the use of gradient descent, which results in a different final recognition saliency range each time. More recognition results are shown in Figure 5.

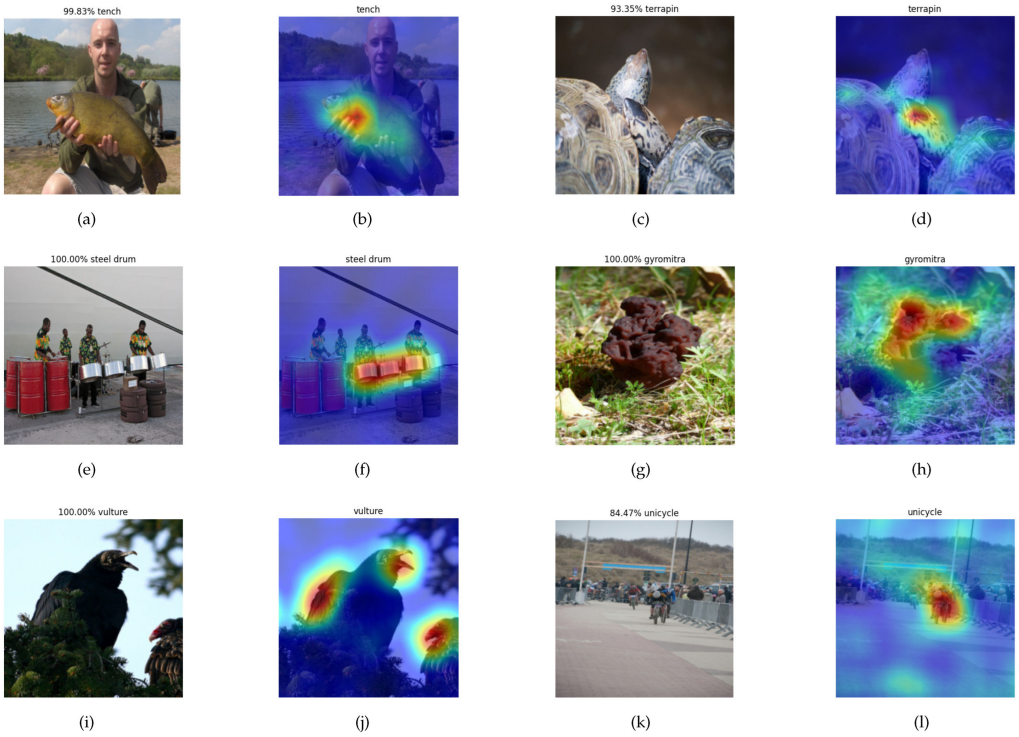


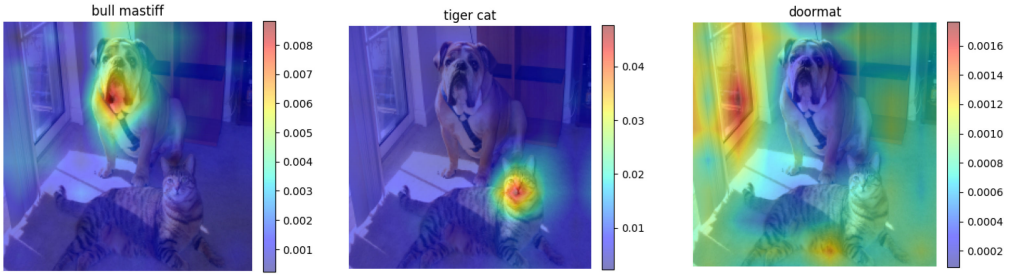
Figure 5. Identification results. (**a**,**c**,**e**,**g**,**i**,**k**) The original input images. (**b**,**d**,**f**,**h**,**j**,**l**) The output images after the OISE algorithm. It can be seen from the figures that the model explains well why these objects are recognized as a tench, terrapin, steel drum, gyromitra, vulture, and unicycle.

4.2. Multi-Target Positioning

Compared to previous methods, this work significantly narrows the significant range of recognition results, allowing for more accurate localization. For example, when identifying an image as a bull mastiff, both the face and the body are responsible for the recognition result, but the facial features are more indicative of the animal being a bull mastiff than the body. It also solves the drawback that the previous method could not or did not perform the classification task better. Figure 6 shows the results and the reasons for OISE to recognize different target objects in the same image. This shows that OISE can classify multiple targets. The color change of the saliency map indicates the different importance of each pixel, with the red color indicating the most important part.

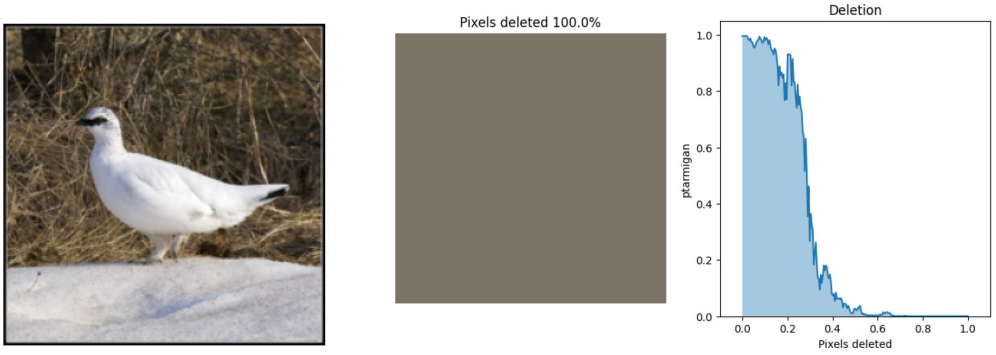
4.3. Evaluation

There are two automatic evaluation metrics: deletion and insertion. Deletion changes the decision of the underlying model. As more important pixels are removed from the image, the probability of predicting a category decreases. The importance of a pixel is defined by its significance score. A sharp drop in the probability curve indicates a good explanation. The insertion metric, on the other hand, uses a complementary approach. As pixels are inserted, the probability of measurement increases, and the higher the area under the curve (AUC), the better the interpretation. When a pixel is removed from an image, the pixel value can be set to 0 or some other constant value. Similarly, pixel insertion can start with a highly blurred image and gradually increase the blurred area. The results are shown in Figure 7.

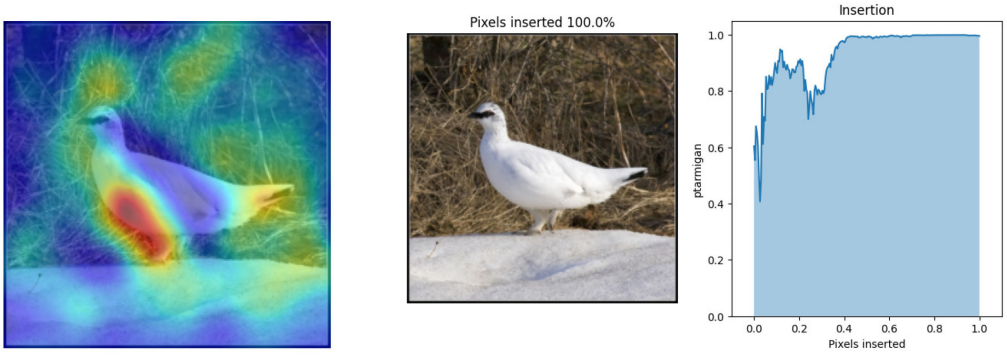


(a) Bull mastiff (b) Tiger cat (c) Doormat

Figure 6. Recognizing multiple targets results in the same image. By visually highlighting the faces of a bull mastiff and a tiger cat, we demonstrate that facial features are key to recognizing differences in animal classes. After pre-training the model, the background is identified as a doormat and highlighted.



(a) Original image (b) Deletion



(c) Saliency map (d) Insertion

Figure 7. Deletion and insertion. (**a**) The original image. (**b**) The result of using the OISE algorithm. (**c**) The image with deletion curves. (**d**) The image with the insertion curves. A sharp drop in the curve when pixels are deleted means a good explanation (the lower the better). When pixels are inserted, a higher AUC indicates a better interpretation (the higher, the better).

4.4. Discussion

The main idea of the algorithm is to summarize the statistics of different features and visualize the significance to establish the causal relationship between features and predictions. Many explainability methods perform summary statistics for each feature based on the decision results and return a quantitative metric, such as feature importance, to measure the importance of different features on the prediction results, and visualize the statistical information of feature importance to visually display the significance graph of important features. The method needs to pre-train a large number of images under classification labels, and the more images and categories are trained, the more accurate the results will be.

5. Conclusions

We propose Optimized Input Sampling Explanation (OISE), a new black-box explanation method for visual interpretation. OISE reduces the computational complexity of randomly generated masks by optimizing the loss function, continuously updating the masks, minimizing the area of the generated masks, and influencing the decision scores, and finally generating saliency maps to show the reasons why the model makes the final decision, thus providing a better explanation to the user. We analyze the evaluation and compare the results, and the method outperforms previous CAM-based methods for better visual explanations in multi-objective classification tasks, making machine decisions more transparent and credible. The method can be applied in many domains, such as the XAI problem for visual target detection, where it can be integrated into a model to generate an interpretation of the target detection, i.e., a bounding box. At the detection level, an attention map is computed to evaluate what information leads to a particular decision.

A good visual explanation will increase people's confidence in the black box model, and with the continuous development of science and technology level, more accurate explainable models can be more widely used in various fields such as medicine, automobiles, and industry to reduce human workload. However, the accuracy of current methods still needs to be improved, which is an important issue that we must continue to explore. In future research on deep learning explainability, we can focus on how to merge different model interpretation techniques to build a more powerful model interpretation method; develop metrics for interpretation methods to measure the interpretation results of models in a more rigorous way; and explore the interpretation work of unsupervised and self-supervised methods to give stronger explainability to models, ensure their fairness, increase privacy protection performance and robustness, and improve users' trust in explainable systems.

Author Contributions: Conceptualization, Z.W.; methodology, Z.W.; software, Z.W.; validation, Z.W. and I.J.; writing—original draft preparation, Z.W.; writing—review and editing, Z.W. and I.J.; visualization, Z.W.; supervision, I.J.; project administration, Z.W. and I.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2020-0-00107, Development of the technology to automate the recommendations for big data analytic models that define data characteristics and problems).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

XAI	Explainable Artificial Intelligence
OISE	Optimized Input Sampling Explanation
AI	Artificial Intelligence
RISE	Randomized Input Sampling for Explanation
CAM	Class Activation Mapping
Grad-CAM	Gradient-weighted Class Activation Mapping
CIC	Channel-wise Increase of Confidence
LIME	Local Interpretable Model Diagnostic Explanations
AUC	Area Under Curve

References

1. Nandhini Abirami, R.; Durai Raj Vincent, P.M.; Srinivasan, K.; Tariq, U.; Chang, C.Y. Deep CNN and deep GAN in computational visual perception-driven image analysis. *Complexity* **2021**, *2021*, 1–30. [CrossRef]
2. Arrieta, A.B.; Diaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
3. Murtaza, G.; Shuib, L.; Abdul Wahab, A.W.; Mujtaba, G.; Mujtaba, G.; Nweke, H.F.; Al-garadi, M.A.; Zulfiqar, F.; Raza, G.; Azmi, N.A. Deep learning-based breast cancer classification through medical imaging modalities: State of the art and research challenges. *Artif. Intell. Rev.* **2020**, *53*, 1655–1720. [CrossRef]
4. Grigorescu, S.; Trasnea, B.; Cocias, T.; Macesanu, G. A survey of deep learning techniques for autonomous driving. *J. Field Robot.* **2020**, *37*, 362–386. [CrossRef]
5. Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Glaeser, C.; Timm, F.; Wiesbeck, W.; Dietmayer, K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 1341–1360. [CrossRef]
6. Gupta, A.; Anpalagan, A.; Guan, L.; Khwaja, A.S. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array* **2021**, *10*, 100057. [CrossRef]
7. Lee, S.M.; Seo, J.B.; Yun, J.; Cho, Y.H.; Vogel-Claussen, J.; Schiebler, M.L.; Gefter, W.B.; Van Beek, E.J.; Goo, J.M.; Lee, K.S.; et al. Deep learning applications in chest radiography and computed tomography. *J. Thorac. Imaging* **2019**, *34*, 75–85. [CrossRef] [PubMed]
8. Torres, A.D.; Yan, H.; Aboutalebi, A.H.; Das, A.; Duan, L.; Rad, P. Patient facial emotion recognition and sentiment analysis using secure cloud with hardware acceleration. In *Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications*; Academic Press: Cambridge, MA, USA , 2018; pp. 61–89.
9. Chen, R.; Yang, L.; Goodison, S.; Sun, Y. Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. *Bioinformatics* **2020**, *36*, 1476–1483. [CrossRef] [PubMed]
10. Öztürk, Ş.; Özkaya, U.; Akdemir, B.; Seyfi, L. Soft tissue sacromas segmentation using optimized otsu thresholding algorithms. *Int. J. Eng. Technol. Manag. Appl. Sci.* **2017**, *5*, 49–54.
11. Şaban, Ö.; Özkaya, U. Skin lesion segmentation with improved convolutional neural network. *J. Digit. Imaging* **2020**, *33*, 958–970.
12. Özkaya, U.; Öztürk, Ş.; Barstugan, M. Coronavirus (COVID-19) classification using deep features fusion and ranking technique. In *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach*; Springer: Cham, Switherland, 2020; pp. 281–295.
13. Sahba, A.; Das, A.; Rad, P.; Jamshidi, M. Image graph production by dense captioning. In Proceedings of the 2018 World Automation Congress (WAC), Stevenson, WA, USA, 3–6 June 2018.
14. Bendre, N.; Ebadi, N.; Prevost, J.J.; Najafirad, P. Human action performance using deep neuro-fuzzy recurrent attention model. *IEEE Access* **2020**, *8*, 57749–57761. [CrossRef]
15. Ozkaya, U.; Öztürk, Ş.; Tuna, K.; Seyfi, L.; Akdemir, B. Faults Detection With Image Processing Methods In Textile Sector. In Proceedings of the 1st International Symposium on Innovative Approaches in Scientific Studies, Padang, Indonesia, 13–14 November 2018.
16. Arun, D.; Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv* **2020**, arXiv:2006.11371.
17. Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; Su, J.K. This looks like that: Deep learning for interpretable image recognition. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.
18. Wang, H.; Wu, X.; Huang, Z.; Xing, E.P. High-frequency component helps explain the generalization of convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
19. Zhang, Q.; Yang, Y.; Ma, H.; Wu, Y.N. Interpreting cnns via decision trees. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–19 June 2019.
20. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–42. [CrossRef]
21. Petsiuk, V.; Das, A.; Saenko, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv* **2018**, arXiv:1806.07421.
22. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
23. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
24. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020.

25. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018.
26. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why should i trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
27. Simonyan, K.; Andrew, Z. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
28. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.